One- and Two-Way Models for Collaborative Studies

Mark G. Vangel
Statistical Engineering Division
National Institute of
Standards and Technology
Building 820, Room 353
Gaithersburg, MD 20899-0001

November 16, 1998

Outline

- **Part 1:** A single material measured by multiple laboratories – *one-way random model* (heteroscedastic and unbalanced)
  - Likelihood Analysis
  - Bayesian Model and Credible Regions
  - Examples

- **Part 2:** Multiple materials meausred by multiple laboratories – *two-way mixed model*
  - Likelihood Analysis
  - Preliminary Work on a Bayesian Model
  - An Example

Part 1: One-Way Model

- The ith of p laboratories makes n_i repeated measurements of the same quantity.

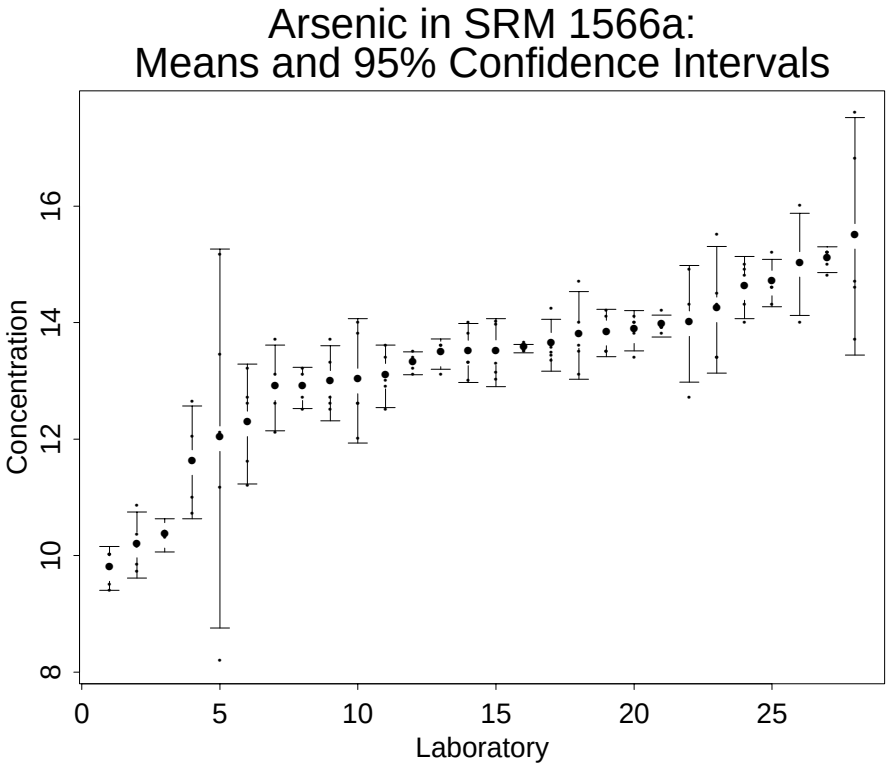
- The laboratories make measurements with different precisions.

- The selected laboratories can be regarded as a random sample from a population (i.e., they're exchangible).

The Problem

How should one estimate the 'grand mean' and between-laboratory variance?

Example #1: Arsenic in Oyster Tissue (NIST Standard Reference Material 1566a)



Example #2: Dietary Fiber
Li and Cardozo (1994)
J. Of AOAC Int., 77, p. 689

Nine labs each measures fiber in *six* foods, in blind duplicates. We will use individual foods for one-way examples, and return to full two-way table later.

Sample	Laboratory			
	1	2	$\cdots$	9
Apples	12.44	12.87	$\cdots$	12.08
	12.48	13.20	$\cdots$	12.38
Apricots	25.05	27.16	$\cdots$	25.31
	25.58	26.29	$\cdots$	25.43
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
FIBRIM	74.07	76.55	$\cdots$	73.96
	75.01	78.36	$\cdots$	74.24

Dietary Fiber in Apricots
Li and Cardozo (1994)

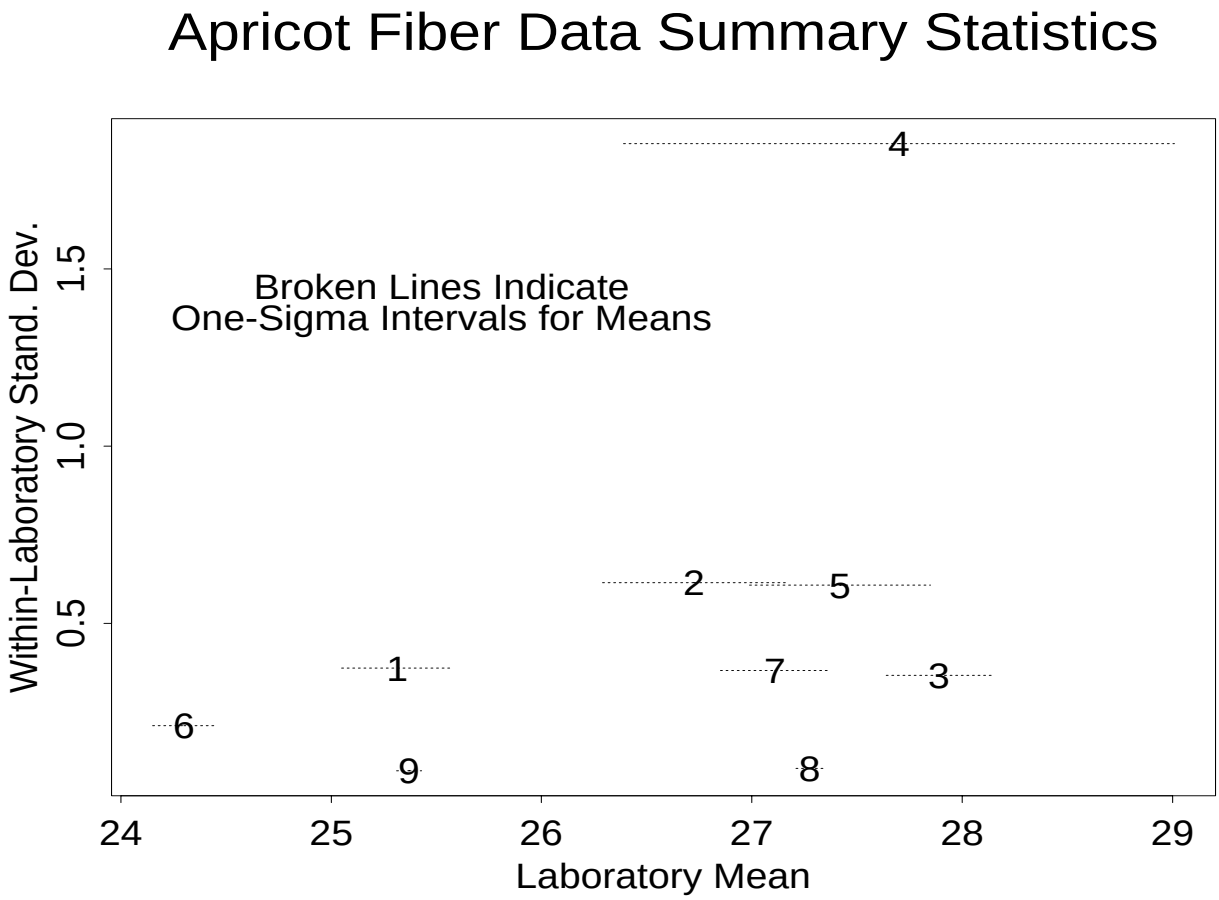
Lab.	x_i	s_i^2	n_i
1	25.32	0.37	2
2	26.72	0.62	2
3	27.89	0.35	2
4	27.70	1.85	2
5	27.42	0.61	2
6	24.30	0.21	2
7	27.11	0.37	2
8	27.28	0.09	2
9	25.37	0.08	2

Mean: $\bar{x} = 26.567$

Weighted Means:

$$\begin{aligned} \text{MP} &= 26.472 \\ \text{GD} &= 26.164 \\ \text{ANOVA} &= 26.420 \\ \text{MLE} &= 27.275 \end{aligned}$$

Plot Of Within-Lab. Standard Deviations vs. Lab. Means for Apricot Fiber Data



Statistical Framework: One-Way, Unbalanced, Heteroscedastic Random-Effects ANOVA

- Laboratory sample means x_i distributed independently normal with mean μ and variance $\sigma^2 + \tau_i^2$, where $\tau_i^2 = \sigma_i^2/n_i$.

- Expected mean for ith laboratory is also normal, with mean μ and variance σ^2.

- Sufficient statistics x_i and $t_i^2 = s_i^2/n_i$.

If x_{ij} denotes the jth measurement from the ith lab, then

$$x_{ij} = \mu + b_i + e_{ij},$$

where $b_i \sim \mathrm{N}(0, \sigma^2)$ and $e_{ij} = \mathrm{N}(0, \sigma_i^2)$; mutually independent.

Cochran's Publications on Combining Experiments

- (1937), "Problems Arising in the Analysis of a Series of Similar Experiments".
- (1938), "The Analysis of Groups of Experiments", (with F. Yates).
- (1954), "The Combination of Estimates From Different Experiments".
- (1980), "Summarizing the Results of a Series of Experiments".
- (1981), "Estimators for the One-Way Random Effects Model With Unequal Error Variances", (et. al., posthumous).

Maximum Likelihood (Cochran, 1937)

Let $\omega_i = 1/(\sigma^2+\tau_i^2)$, $\nu_i = n_i - 1$, and determine $\widehat{\sigma}$, $\widehat{\tau}_i^2$, and $\widehat{\mu}$ to satisfy

$$(A_i) \quad \omega_i - \omega_i^2(x_i-\mu)^2 + \nu_i\left(\frac{1}{\tau_i^2} - \frac{t_i^2}{\tau_i^4}\right) = 0$$

$$(B) \qquad \boxed{\sum_{i=1}^k \omega_i^2(x_i-\mu)^2 = \sum_{i=1}^k \omega_i}$$

$$(C) \qquad \mu = \frac{\sum_{i=1}^k \omega_i x_i}{\sum_{i=1}^k \omega_i}$$

Note that (B) may have multiple roots. Cochran (1937) proposed setting $\tau_i^2 = t_i^2$ and solving (B) for σ^2, then using (C).

The Loglikelihood Function: A Better Parametrization

Define weights by

$$\gamma_i \equiv \frac{\sigma^2}{\sigma^2 + \tau_i^2}$$

The loglikelihood becomes

$$\begin{aligned} 2\ell &= \sum_{i=1}^{p} n_i \log\left(\frac{\gamma_i}{\sigma^2}\right) \\ &- \sum_{i=1}^{p} \frac{\gamma_i}{\sigma^2}\left[(x_i - \mu)^2 + \frac{\nu_i t_i^2}{1-\gamma_i}\right] \\ &- \sum_{i=1}^{p} \nu_i \log(1-\gamma_i) + K. \end{aligned}$$

Differentiate this with respect to parameters μ, σ^2 and $\gamma_i, i = 1, \ldots, p$.

ML Equations

$$\mu = \frac{\sum_{i=1}^{p} \gamma_i x_i}{\sum_i \gamma_i} = \frac{\sum_{i=1}^{p} \omega_i x_i}{\sum_i \omega_i}$$

$$\sigma^2 = \frac{\sum_{i=1}^{p} \gamma_i \left[(x_i - \mu)^2 + \frac{\nu_i t_i^2}{1-\gamma_i} \right]}{\sum_{i=1}^{p} n_i}$$

$$\begin{aligned} &\gamma_i^3 - (a_i + 2)\gamma_i^2 + \\ &[(n_i + 1)a_i + (n_i - 1)b_i + 1]\,\gamma_i \\ &-n_i a_i = 0 \end{aligned}$$

where

$$a_i \equiv \frac{\sigma^2}{(x_i - \mu)^2}$$

and

$$b_i \equiv \frac{t_i^2}{(x_i - \mu)^2}.$$

Result #1:
Monotone Convergence to Stationary Points of the Likelihood

- For any starting values μ_0, σ_0^2, maximize the likelihood over the weights by solving the cubics. (If there are multiple real roots, choose the one which causes the biggest increase in the likelihood.)

- Let

$$\sigma_1^2 = \frac{\sum_{i=1}^{p} \gamma_i \left[(x_i - \mu)^2 + \frac{\nu_i t_i^2}{1-\gamma_i} \right]}{\sum_{i=1}^{p} n_i}$$

$$\mu_1 = \frac{\sum_{i=1}^{p} \gamma_i x_i}{\sum_{i=1}^{p} \gamma_i}$$

  solve for new weights, and iterate.

- This iteration, *regardless of starting values*, always converges to a stationary point of the likelihood, and *increases the likelihood at each step.*

Result #2: Location of Stationary Values of the Likelihood

- At a stationary point of the likelihood,

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{p} \gamma_i^2 (x_i - \mu)^2}{\sum_{i=1}^{p} \gamma_i}$$

  hence

- *All* of the stationary points of the likelihood $\widehat{\mu}$ and $\widehat{\sigma}$ are within the rectangle in the (μ, σ) plane given by

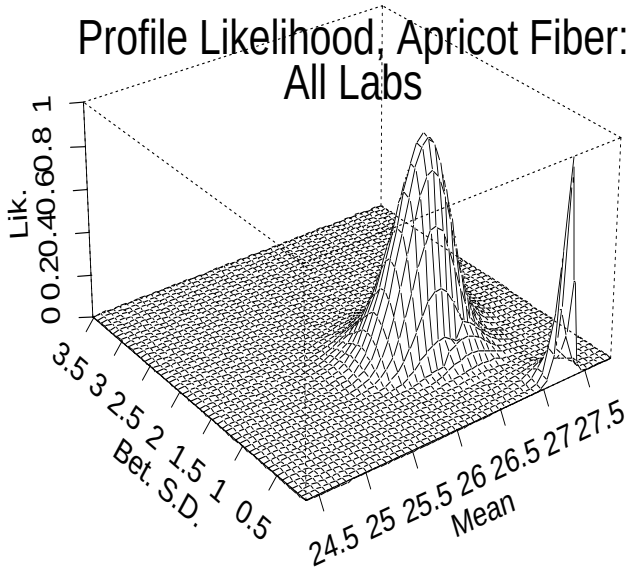
$$\min_i(x_i) \leq \tilde{\mu} \leq \max_i(x_i)$$

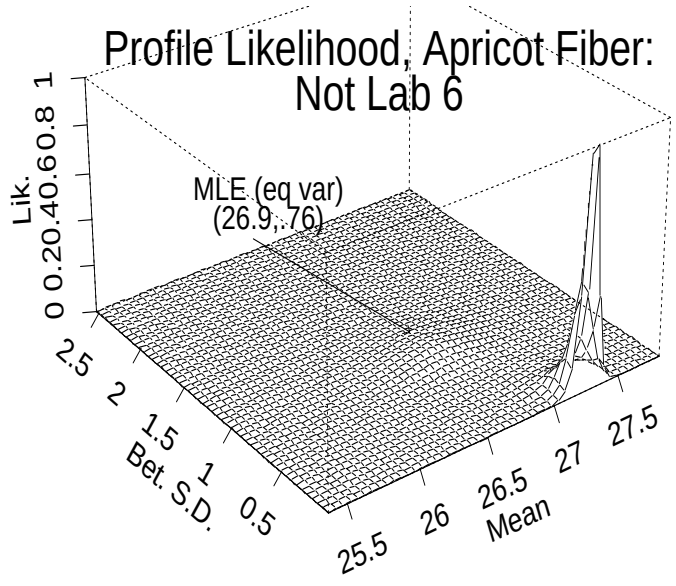
  and

$$0 \leq \tilde{\sigma} \leq \max_i(x_i) - \min_i(x_i).$$

- After the appropriate location-scale transformation of the data, it is only necessary to search the unit square in the (μ, σ) plane for stationary values.

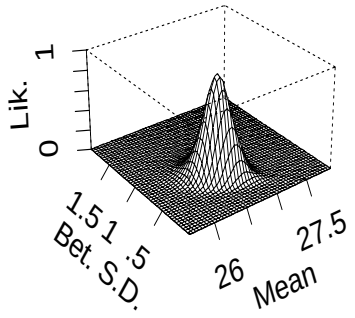
Lab. 6 an Outlier for Apricot Data



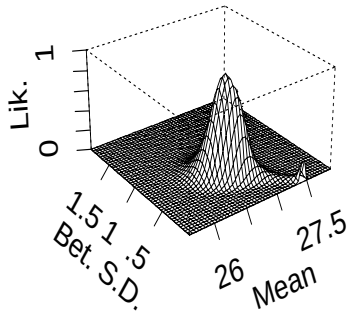


Outlier Labs. for Cabbage Data

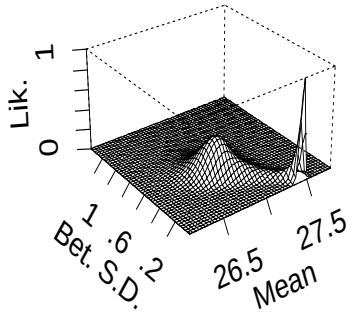
Prof. Lik., Cabbage Fiber:
All Labs



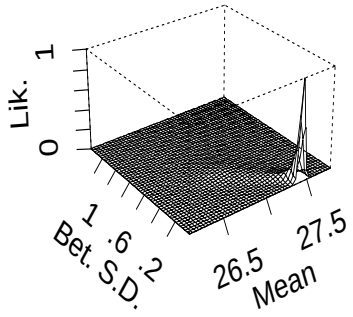
Prof. Lik., Cabbage Fiber:
Not Lab 6



Prof. Lik., Cabbage Fiber:
Not Labs 6, 9



Prof. Lik., Cabbage Fiber:
Not Labs 6, 9, 1



Result #3: Location of the Roots of Cubic Equations for Weights (γ_i)

- Each cubic likelihood equation has one or three roots $\gamma_i \in [0, 1]$.

- A necessary condition for three roots is that

$$(x_i - \mu)^2 \geq \max(\sigma^2/q_i, t_i^2/h_i),$$

where

$$\begin{aligned} q_i &= -2 - 6\sqrt{n_i}\sin\left\{\frac{1}{3}\left[\sin^{-1}\left(\sqrt{\frac{n_i - 1}{n_i}}\right) - \frac{\pi}{2}\right]\right\} \\ &= \frac{8}{27n_i} + O(n_i^{-2}) \end{aligned}$$

and

$$h_i = \frac{(1 - q_i)^3}{27(n_i - 1)} = \frac{1}{27n_i} + O(n_i^{-2}).$$

- These values q_i and h_i are the smallest for which this is necessary.

Outline of Proof of Result #3 (The subscript i is omitted for clarity.)

- Descartes' rule of signs applied to the original equation and to

$$\phi^3 + (1-a)\phi^2 + \nu(a+b)\phi + \nu b = 0,$$

where $\phi = \gamma - 1$ shows that

- the number of roots in $[0, 1]$ is 1 or 3.

Proof of #3, Continued

- Assume three roots r_j of 'γ' equation and roots $d_j = r_j - 1$ of 'ϕ' equation in $[0, 1]$.

- Quadratic coefficient of a monic cubic equals negative of sum of roots, and constant term equal negative of product of roots.

- From 'γ' equation

$$\left(\frac{a+2}{3}\right)^3 \geq na, \text{ or}$$

$$f(a) = a^3 + 6a^2 + (12 - 27n)a + 8 \geq 0,$$

where $f(a)$ has one negative root, one greater than 1, and one (q) in $(0, 1/n)$, given earlier.

- From the 'ϕ' equation:

$$\left(\frac{1-a}{3}\right)^3 \geq (n-1)b.$$

When evaluated at q, this completes the derivation of the necessary condition.

Proof of #3, Concluded

The region in the (a, b) plane which corresponds to three real roots is bounded by the locus of (a, b) pairs for which the discriminant is zero, and as a consequence two or three of these real roots coincide. At the upper-right-hand corner of the rectangle $[0, q] \times [0, h]$ the two inequalities on the previous transparency become equalities, the three real roots are equal, and the discriminant is zero. Any smaller rectangle must exclude (a, b) pairs for which the discriminant is positive, and hence for which there will be three roots in [0,1].

A Comment on Homoscedastic Models

- If we require that the 'within' variances be equal, than we still have the likelihood equations

$$\mu = \frac{\sum_{i=1}^{p} \gamma_i x_i}{\sum_i \gamma_i}$$

$$\sigma^2 = \frac{\sum_{i=1}^{p} \gamma_i \left[(x_i - \mu)^2 + \frac{\nu_i t_i^2}{1-\gamma_i} \right]}{\sum_{i=1}^{p} n_i}$$

- The weights can be parametrized as

$$\gamma_i = \frac{u}{u + (1-u)\frac{n_1}{n_i}},$$

for $0 \leq u < 1$. Maximizing the likelihood reduces to maximizing with respect to u. In particular, all of the stationary points of the likelihood must be on a curve in the (μ, σ) plane, and one need not be concerned about negative solutions for the variances.

Hierarchical Model With Noninformative Priors

$i = 1, \ldots, p$ indexes laboratories

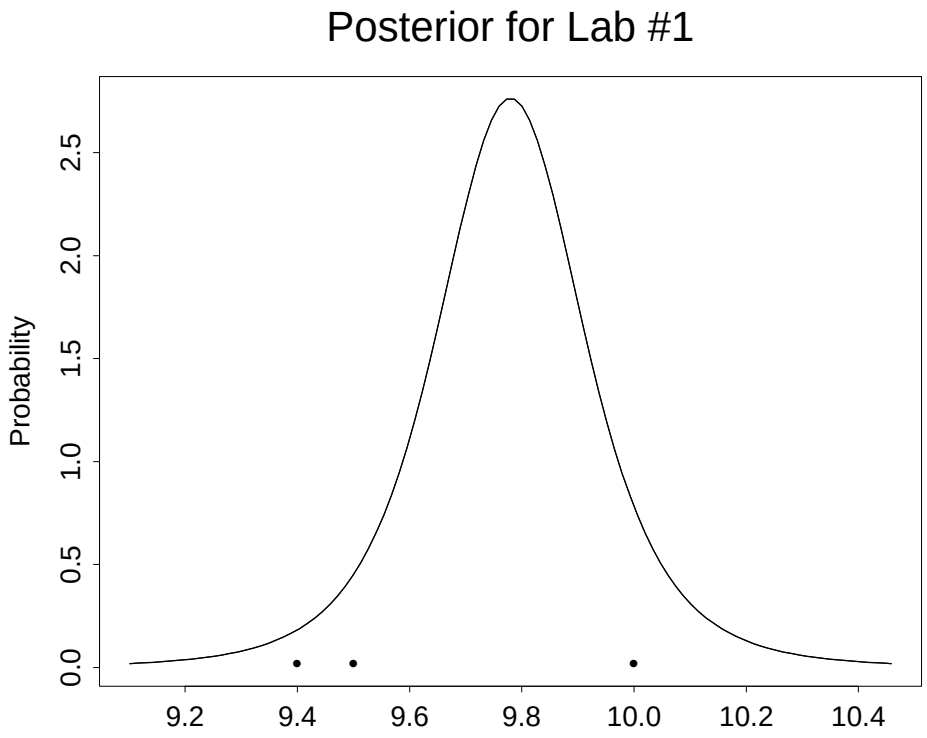
$j = 1, \ldots, n_i$ indexes measurements

$$\begin{aligned} p(x_{ij}|\delta_i, \sigma_i^2) &= \mathrm{N}(\delta_i, \sigma_i^2) \\ p(\sigma_i) &\propto 1/\sigma_i \\ p(\delta_i|\mu, \sigma^2) &= \mathrm{N}(\mu, \sigma^2) \\ p(\mu) &= 1 \\ p(\sigma) &= 1 \end{aligned}$$

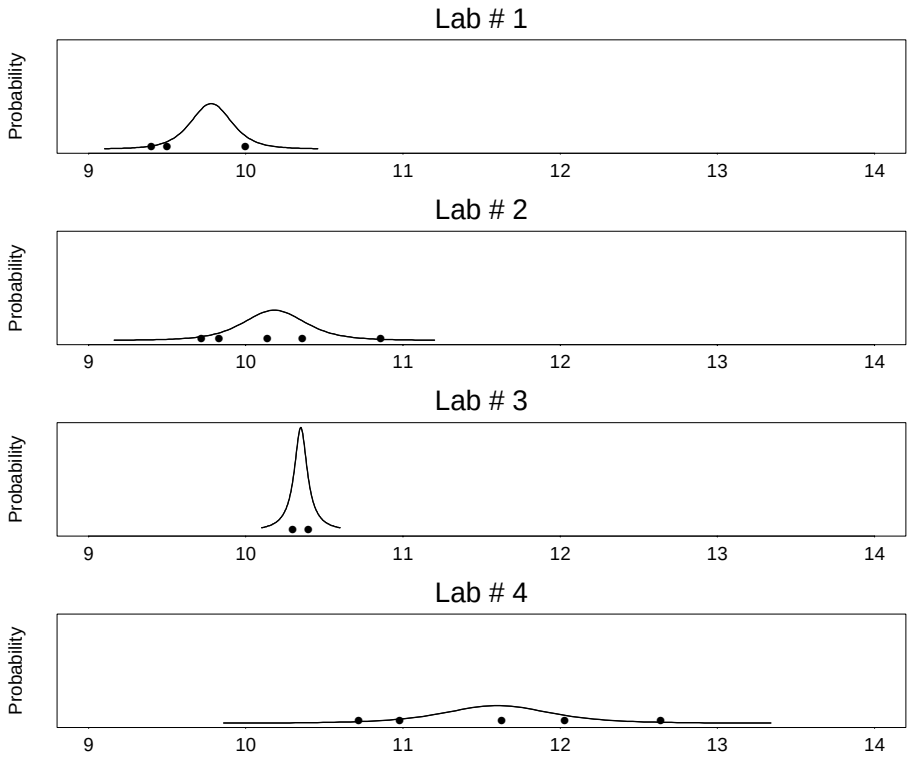
Arsenic Data Posterior for μ: $\sigma = 0$ and $p = 1$

If $\sigma = 0$ and $p = 1$, then

$$p(\mu|\{x_i\}, \sigma = 0) = \frac{\sqrt{n}}{s} T'_{n-1}\left(\frac{\mu - \bar{x}}{s/\sqrt{n}}\right)$$



Arsenic Data Posteriors for the First Four Labs, Given $\sigma = 0$



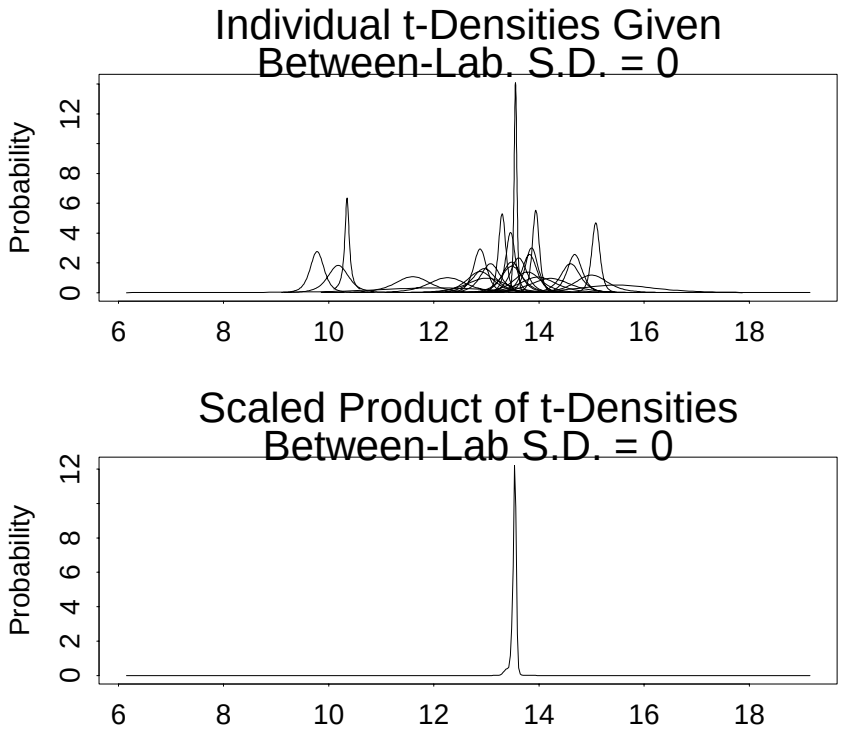
Posterior for $\sigma = 0$, $p \geq 1$

Given $\sigma = 0$, then the posterior distribution of the consensus mean μ is proportional to a product of scaled t-densities:

$$p(\mu|\{x_{ij}\}, \sigma = 0) \propto \prod_{i=1}^{p} \frac{1}{t_i} T'_{n_i - 1}\left(\frac{\bar{x}_i - \mu}{t_i}\right)$$

Posterior of μ for Arsenic Data Given $\sigma = 0$

The posterior is proportional to the *product* of the appropriate t-densities, centered at each lab average $\bar{x}_i$.



The General Case: $\sigma \geq 0$

In general, $p(\mu|\sigma, \{x_{ij}\})$ is proportional to a *product* of the distributions of the random variables

$$U_i = \bar{x}_i + \frac{\sqrt{n_i}}{s_i} T_{n_i-1} + \sigma Z,$$

where T_{n_i-1} is a t-distributed random variable with $n_i - 1$ degrees of freedom, Z is distributed $\mathsf{N}(0, 1)$, and T_{n_i-1} and Z are independent.

A Useful Probability Density

Let T_ν and Z denote independent Student-t and standard normal random variables, and assume that $\psi \geq 0$ and $\nu > 0$. Then

$$U = T_\nu + Z\sqrt{\frac{\psi}{2}}$$

has density

$$f_\nu(u;\psi) \equiv \frac{1}{,\ _{\nu/2}\sqrt{\pi}} \int_0^\infty \frac{y^{(\nu+1)/2-1} e^{-y\left[1+\frac{u^2}{\psi y+\nu}\right]}}{\sqrt{\psi y+\nu}} dy.$$

Posterior of (μ, σ)

- Assume $\delta_i \sim \mathrm{N}(\mu, \sigma^2)$, $\sigma \sim p(\sigma)$, $p(\mu) = 1$, $p(\sigma_i) = 1/\sigma_i$.

- Then the posterior of (μ, σ) is

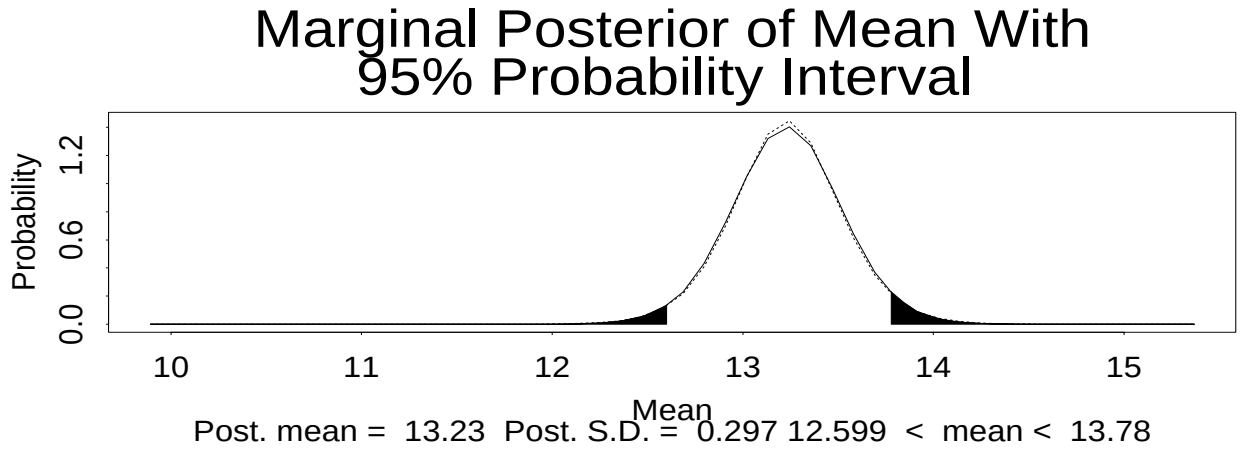
$$p(\mu, \sigma | \{x_{ij}\}) \propto p(\sigma) \prod_{i=1}^{p} \frac{1}{t_i} f_{n_i - 1}\left[\frac{x_i - \mu}{t_i}; \frac{2\sigma^2}{t_i^2}\right].$$

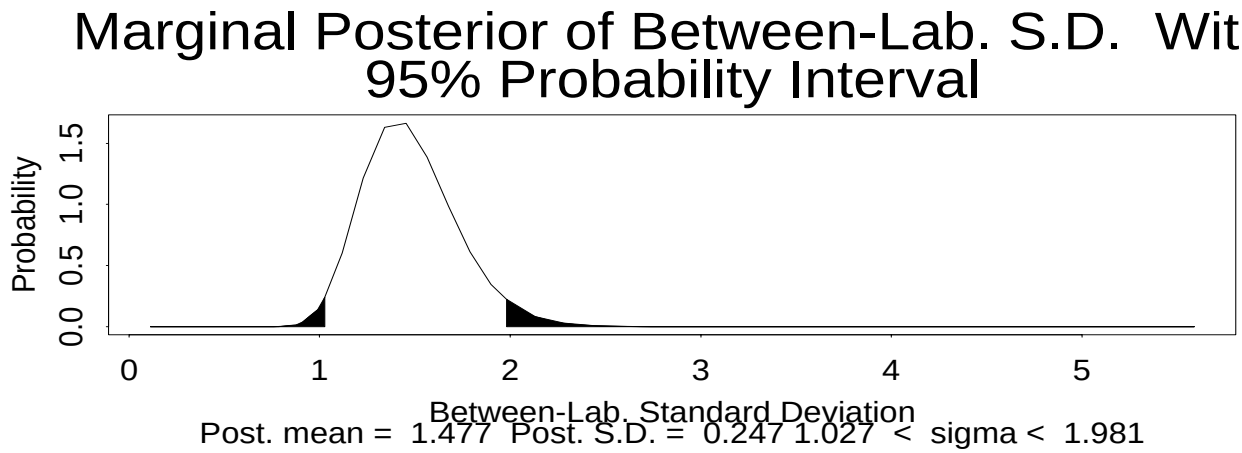
- The posterior of μ given $\sigma = 0$ is a product of scaled t-densities centered at the x_i, since

$$\frac{1}{t_i} f_{n_i - 1}\left[\frac{x_i - \mu}{t_i}; 0\right] = \frac{1}{t_i} T'_{n_i - 1}\left(\frac{x_i - \mu}{t_i}\right).$$

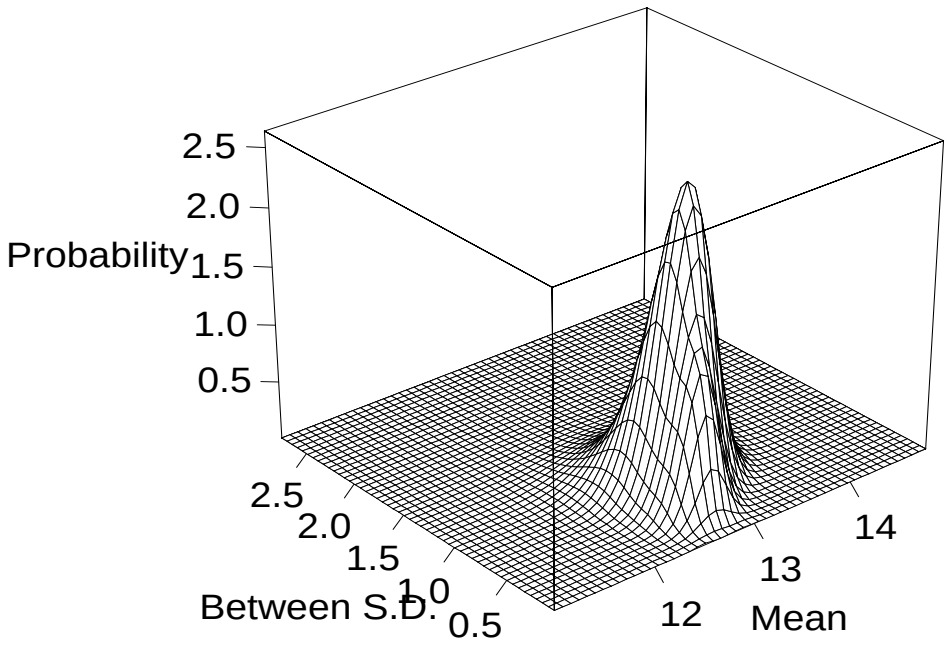
- We will take $p(\sigma) = 1$, though an arbitrary proper prior does not introduce additional difficulties.

Marginals for μ and σ: Arsenic Data

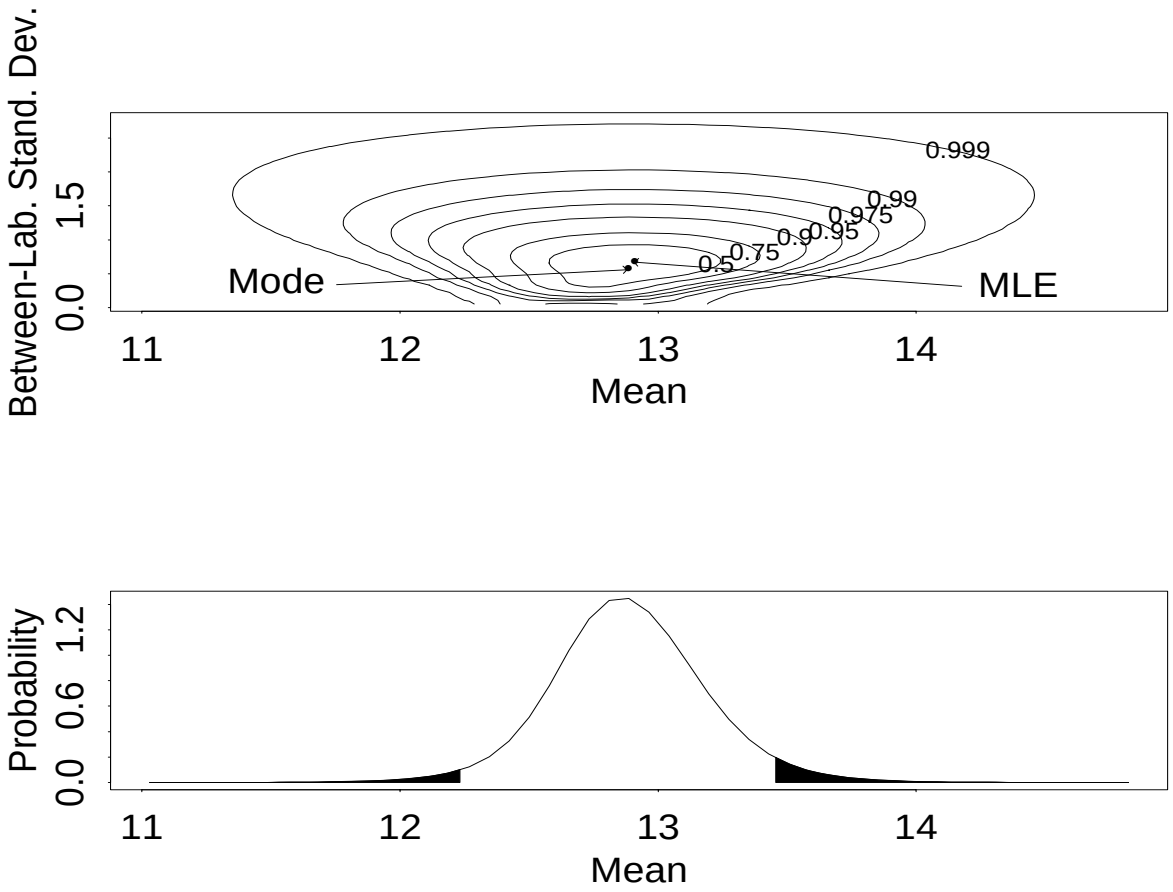




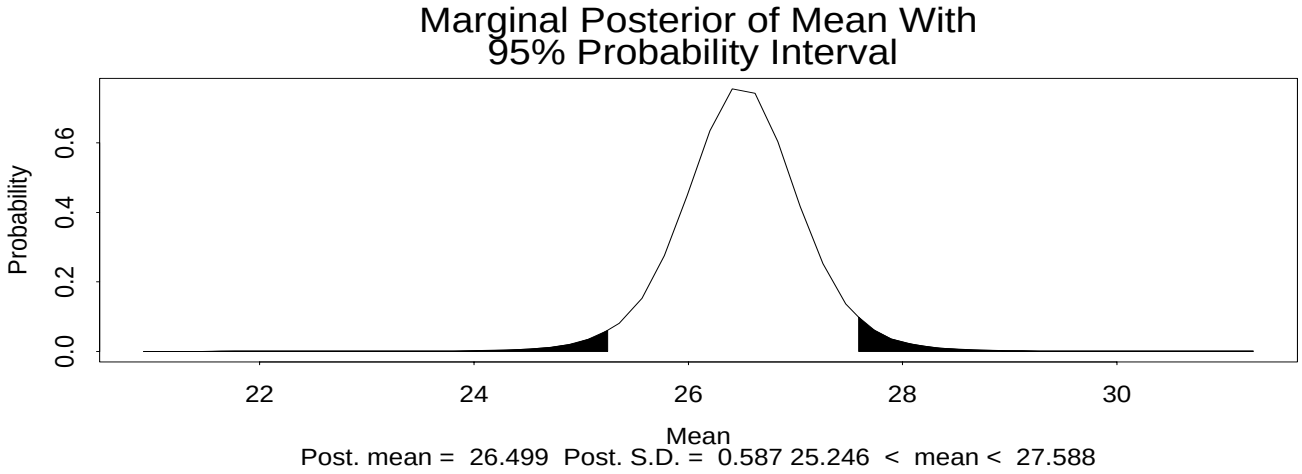
Marginal Posterior: Apple Fiber Data

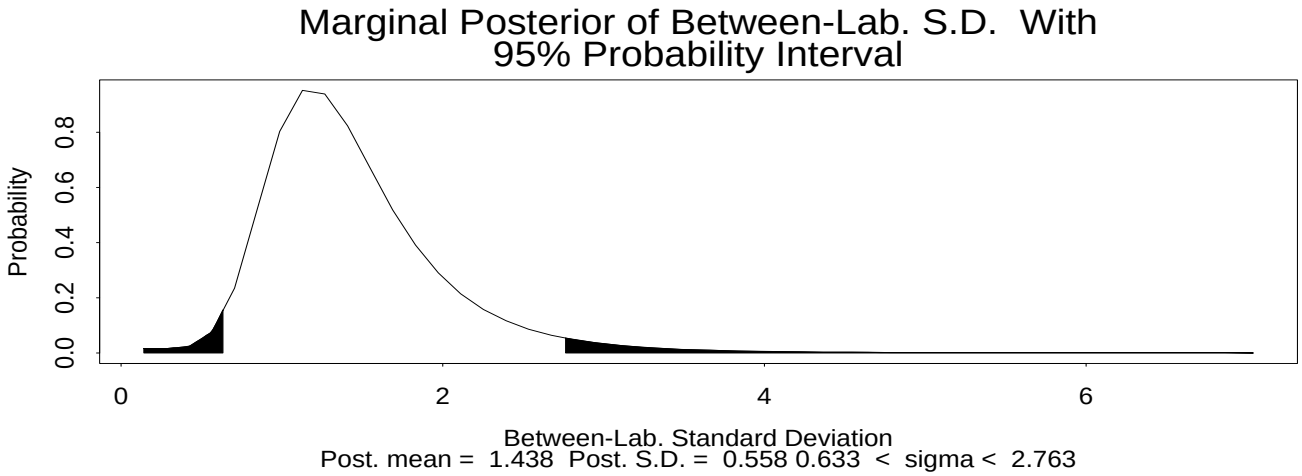


Approximate Confidence Regions: Apple Fiber Data



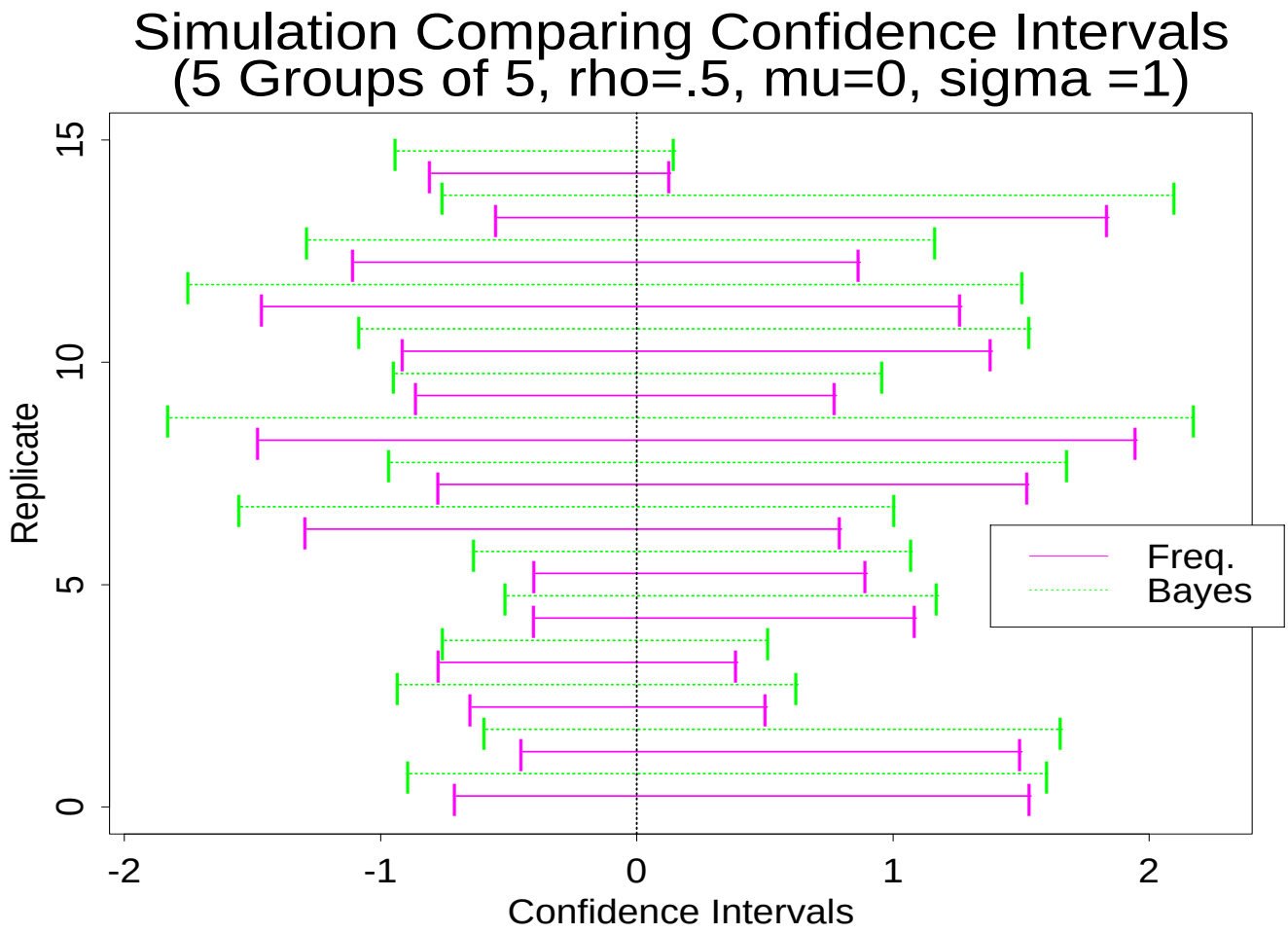
Approximate Confidence Intervals: Apricot Fiber Data





Small Simulation Comparing Bayesian and Frequentist Intervals

$$\begin{aligned} \mu &= 0 \\ \sigma_i &= \sigma_e \\ \sigma^2 + \sigma_e^2 &= 1 \\ \rho &= \sigma^2/(\sigma_e^2 + \sigma^2) = 1/2 \end{aligned}$$

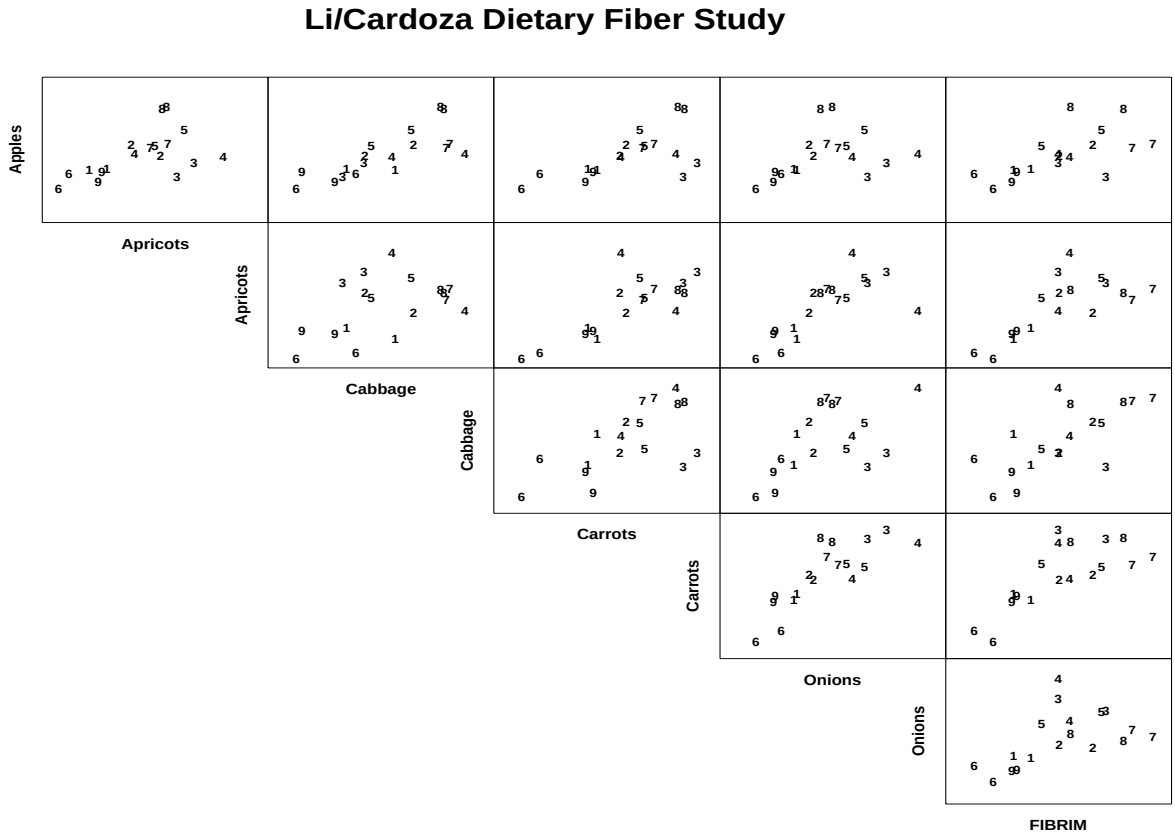


Part 2: Two-Way Mixed Model Multiple Materials and Multiple Laboratories

- Each of p laboratories makes repeated measurements of m materials.
- The number of measurements made can differ among the laboratories, but each material is measured the same number of times by each laboratory.
- The within-laboratory variances can differ.
- The selected laboratories can be regarded as a random sample from an infinite population of laboratories.

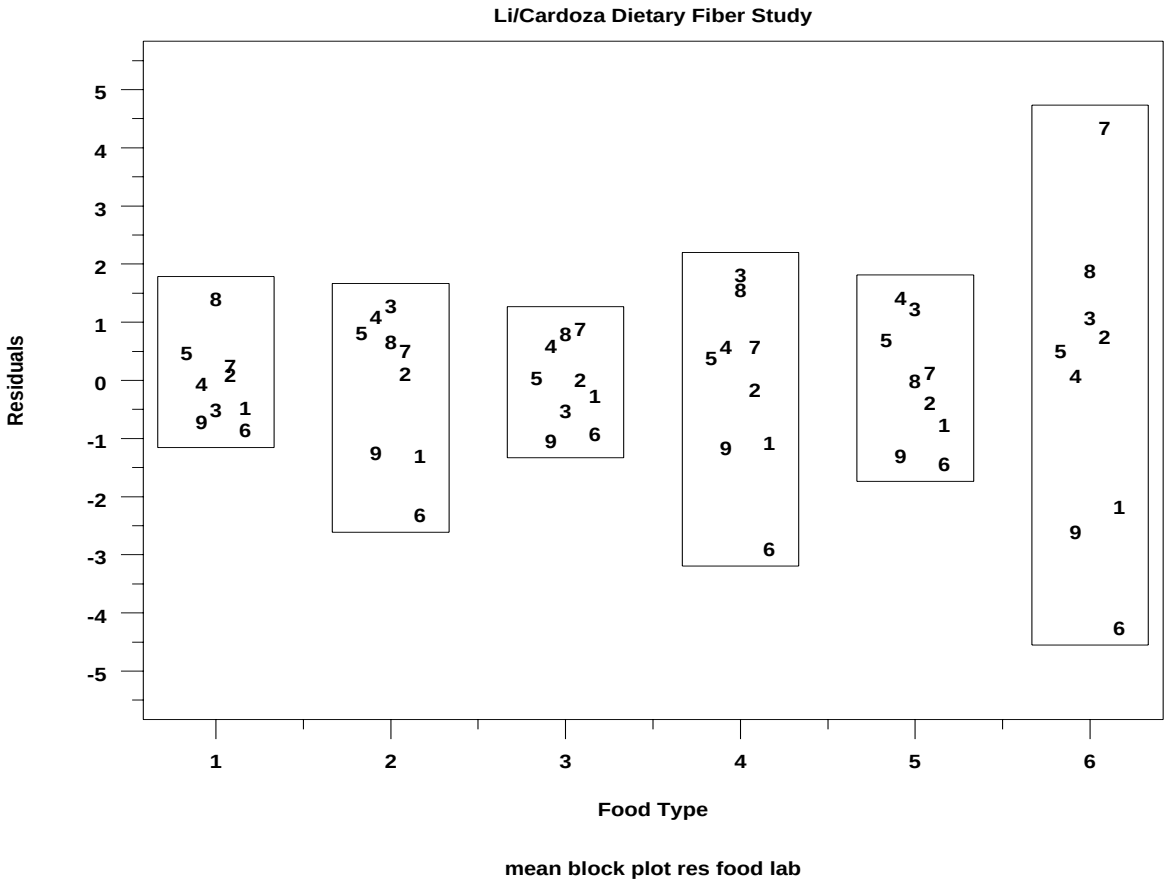
How should one estimate 'consensus' values for the quantities measured, and what are the uncertainty in this estimates?

Dietary Fiber: 9 Labs, 6 Foods, 2 Replicates



vangel41

Block Plot of Fiber Data



mean block plot res food lab

vangel42.

Two-Way Mixed Model (Heteroscedastic no Interaction)

$$x_{ijk} = \theta_k + \delta_i + e_{ijk},$$

- $i = 1, \ldots, p$ Laboratories
- $j = 1, \ldots, n_i$ Replicates
- $k = 1, \ldots, m$ Materials

$$\delta_i \sim \mathsf{N}(0, \sigma^2)$$

$$e_{ijk} \sim \mathsf{N}(0, \sigma_i^2)$$

Some notation: $\tau_i^2 \equiv \sigma_i^2/(n_i m)$, $\nu_i \equiv n_i m - 1$.

Two-Way Model: Likelihood

Model:

$$x_{ijk} = \theta_k + \delta_i + e_{ijk},$$

Likelihood:

$$L \propto \prod_{i=1}^{p} \int_{-\infty}^{\infty} \frac{e^{-\delta_i^2/2\sigma^2}}{\sqrt{\sigma^2(\sigma_i^2)^{n_i m}}} \prod_{j=1}^{n_i} \prod_{k=1}^{m} e^{-(x_{ijk}-\theta_k-\delta_i)^2/2\sigma_i^2} \, \mathrm{d}\delta_i$$

$$\propto \prod_i \frac{\exp\left[-\frac{(\bar{x}_{i..}-\bar{\theta})^2}{2(\sigma^2+\tau_i^2)}\right]}{\sqrt{(\tau_i^2)^{\nu_i}(\sigma^2+\tau_i^2)}}$$

$$\cdot \exp\left[-\frac{\sum_{j,k}(x_{ijk}-\bar{x}_{i..}+\bar{\theta}-\theta_k)^2}{2n_i m\tau_i^2}\right]$$

ML Equations

$$\theta_k - \bar{\theta} = \frac{\sum_{i=1}^{p} (\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot})/\tau_i^2}{\sum_{i=1}^{p} 1/\tau_i^2}$$

$$\bar{\theta} = \frac{\sum_{i=1}^{p} \gamma_i \bar{x}_{i\cdot\cdot}}{\sum_{i=1}^{p} \gamma_i}$$

$$\sigma^2 = \frac{\sum_{i=1}^{p} \gamma_i \left[(x_i - \mu)^2 + \frac{\nu_i t_i^2}{1-\gamma_i} \right]}{\sum_{i=1}^{p} n_i}$$

Where $\tau_i^2 \equiv \sigma_i^2/(n_i m)$, $\nu_i \equiv m n_i - 1$,
$\gamma_i \equiv \sigma^2/(\sigma^2 + \tau_i^2)$, and

$$t_i^2 \equiv \frac{\sum_{j,k} (x_{ijk} - \bar{x}_{i\cdot k})^2 + n_i \sum_k (\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot} + \bar{\theta} - \theta_k)^2}{\nu_i n_i m}$$

ML Equations (Cont'd)

The weights $\{\gamma_i\}_{i=1}^{p}$ are roots of the cubic equations

$$\gamma_i^3 - (a_i + 2)\gamma_i^2 + [(n_i m + 1)a_i + \nu_i b_i + 1]\,\gamma_i - n_i a_i = 0$$

where

$$a_i \equiv \frac{\sigma^2}{(\bar{x}_{i\cdot\cdot} - \bar{\theta})^2}$$

and

$$b_i \equiv \frac{t_i^2}{(\bar{x}_{i\cdot\cdot} - \bar{\theta})^2}.$$

An ML Iteration

1. Begin with estimates $\left\{\gamma_i^{(s)}\right\}$.

2. Calculate the following:

$$\begin{aligned}
\phi_k^{(s+1)} &= \frac{\sum_{i=1}^{p}(\bar{x}_{i\cdot k}-\bar{x}_{i\cdot\cdot})/\tau_i^{2(s)}}{\sum_{i=1}^{p} 1/\tau_i^{2(s)}} \\
\bar{\theta}^{(s+1)} &= \frac{\sum_{i=1}^{p}\gamma_i^{(s)}\bar{x}_{i\cdot\cdot}}{\sum_{i=1}^{p}\gamma_i^{(s)}} \\
\sigma^2_{(s+1)} &= \frac{\sum_{i=1}^{p}\gamma_i^{(s)}\left[(x_i-\mu)^2+\frac{\nu_i t_i^2}{1-\gamma_i^{(s)}}\right]}{\sum_{i=1}^{p} n_i}
\end{aligned}$$

3. Note that if the ϕ_k are constrained to satisfy the above ML equation, then

$$t_i^2 = \frac{\sum_{j,k}(x_{ijk}-\bar{x}_{i\cdot\cdot})^2 - \sum_k \phi_k^2/m}{n_i\nu_i m}$$

4. Solve the cubics for new estimates $\gamma_i^{(s+1)}$, and iterate.

Some Theoretical Results for Two-Way Mixed Model

The one-way results discussed earlier generalize:

- Monotone convergence
- All stationary values of likelihood in box in $(\mu, \sigma, \sum_k \phi_k^2)$ space.
- Exactly one weight $\gamma_i \in [0, 1]$, unless ith lab an outlier and n_i small
- Variances cannot be negative at solution to likelihood equation.

Hierarchical Model With Noninformative Priors: Two-Way Model

$i = 1, \ldots, p$ indexes laboratories

$j = 1, \ldots, n_i$ indexes measurements

$k = 1, \ldots, m$ indexes materials

$$\begin{aligned} p(x_{ijk}|\delta_i, \theta_k, \sigma_i^2) &= \mathsf{N}(\delta_i + \theta_k, \sigma_i^2) \\ p(\sigma_i) &\propto 1/\sigma_i \\ p(\delta_i|\mu, \sigma^2) &= \mathsf{N}(\mu, \sigma^2) \\ p(\theta_k) &= 1 \\ p(\sigma) &= 1 \end{aligned}$$

Posterior of (μ, σ): Two-Way Model

- The posterior of $(\{\theta_k\}, \sigma)$ is

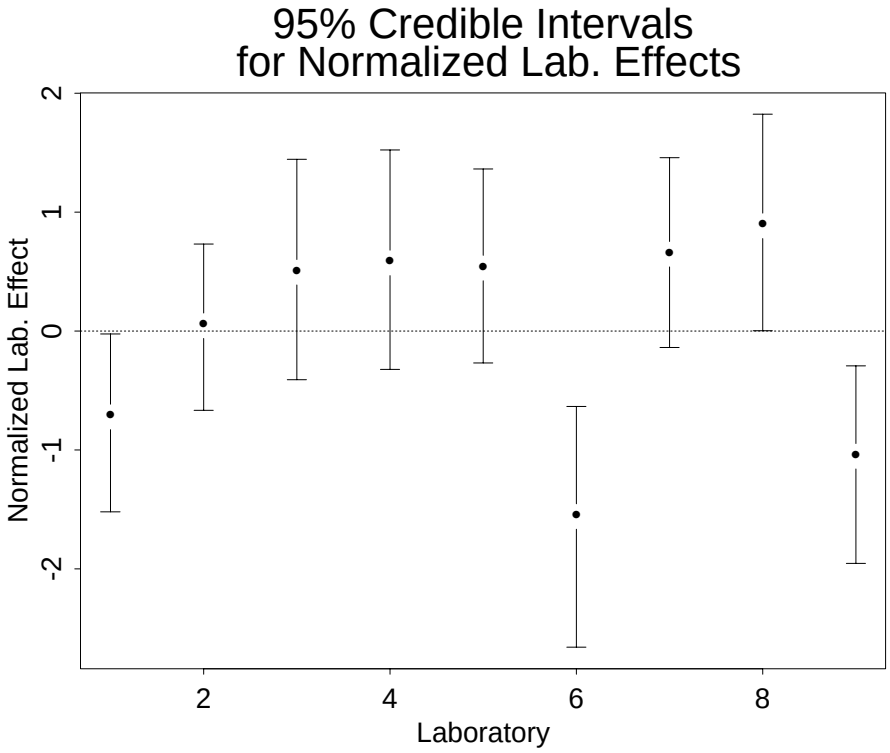
  $$p(\{\theta_k\}, \sigma | \{x_{ijk}\}) \propto p(\sigma) \prod_{i=1}^{p} \frac{1}{t_i} f_{\nu_i} \left[\frac{x_i - \mu}{t_i}; \frac{2\sigma^2}{t_i^2} \right].$$

  where $f_\nu(\cdot, \theta)$ is the and

  $$t_i^2 \equiv \frac{\sum_{j,k}(x_{ijk} - \bar{x}_{i\cdot k})^2 + n_i \sum_k (\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot} + \bar{\theta} - \theta_k)^2}{\nu_i n_i m}$$

- We will take $p(\sigma) = 1$.

An Example Posterior Calculation (BUGS): Credible Intervals for $\{\delta_i/\sigma\}$



Summary

- Using a simple parametrization, ML (and REML) calculations for one- and two-way heteroscedastic models are straightforward. All stationary values of the likelihood can be determined.

- Investigation of multi-modal likelihoods can (at least for the one-way model) lead to useful insight into the data.

- Bayesian calculations by numerical integration are straighforward for the one-way model.

- Work on the two-way table is in progress.